Monte Carlo Methods

Spring Semester 2013/14, Department of Applied Mathematics, University of Crete

Instructor: Harmandaris Vagelis, email: vagelis@tem.uoc.gr

Part III: Sampling Methods

Basic Idea

What we have seen . . .

How to generate uniform $\mathsf{U}[0,1]$ pseudo-random numbers.

This lecture will cover . . .

Generating random numbers from any distribution using

- transformations (CDF inverse, Box-Muller method).
- rejection sampling.

❑ Transformation Methods:

- We can generate
$$U \sim \mathsf{U}[0,1].$$
- Can we find a transformation T such that
$$T(U) \sim F$$
for a distribution of interest with CDF F?
- One answer to this question: inversion method.

Transformation Methods

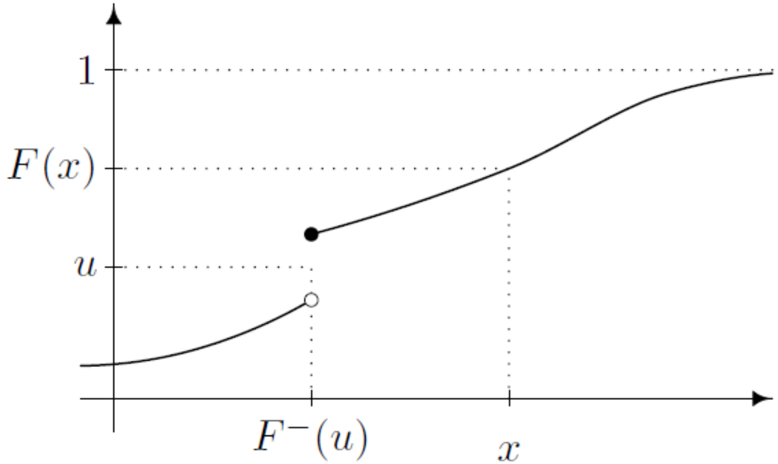
❑ **CDF and its Generalized Inverse:**

Cumulative distribution function (CDF)

$$F(x) = \mathbb{P}(X \leq x)$$

Generalised inverse of the CDF

$$F^{-}(u) := \inf\{x : \ F(x) \geq u\}$$



Properties of F^{-} (taken without proof)

1. $F^{-}(F(x)) \leq x, \quad \forall x \in F^{-}([0,1])$
2. $F(F^{-}(u)) \geq u, \quad \forall u \in [0,1]$

Transformation Methods

❑ **Inversion Method:**

Theorem 2.1: Inversion method

Let $U \sim \mathrm{U}[0,1]$ and F be a CDF. Then $F^{-}(U)$ has the CDF F.

Proof: From the definition of the CDF, $F(x) = \mathbb{P}(U \le F(x))$, so we need to prove that

$$\mathbb{P}(F^{-}(U) \le x) = \mathbb{P}(U \le F(x)), \quad \forall x.$$

It is sufficient to prove the equivalence:

$$F^{-}(U) \le x \Leftrightarrow U \le F(x).$$

Inverse Method

- **Example: Exponential Distribution**

The exponential distribution with rate $\lambda > 0$ has the CDF ($x \geq 0$)

$$\begin{aligned} F_\lambda(x) &= 1 - \exp(-\lambda x) \\ F_\lambda^-(u) &= F_\lambda^{-1}(u) = -\log(1-u)/\lambda. \end{aligned}$$

So we have a simple algorithm for drawing Expo(λ):

1. Draw $U \sim \mathrm{U}[0,1]$.
2. Set $X = -\frac{\log(1-U)}{\lambda}$, or equivalently $X = -\frac{\log(U)}{\lambda}$.

Inverse Method

❑ **Example: Box - Muller method for Generating Gaussians**

Box-Muller method

1. Draw
$$U_1, U_2 \overset{\text{i.i.d.}}{\sim} \mathsf{U}[0,1].$$
2. Set
$$\begin{aligned} X_1 &= \sqrt{-2\log(U_1)} \cdot \cos(2\pi U_2), \\ X_2 &= \sqrt{-2\log(U_1)} \cdot \sin(2\pi U_2). \end{aligned}$$

Then $X_1, X_2 \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0,1)$.

Inverse Method

❑ Example: Box - Muller method for Generating Gaussians

- Consider a bivariate real-valued random variable (X_1, X_2) and its polar coordinates (R, θ), i.e.

$$X_1 = R \cdot \cos(\theta), \qquad X_2 = R \cdot \sin(\theta) \qquad (1)$$

- Then the following equivalence holds:
$X_1, X_2 \overset{\text{i.i.d.}}{\sim} \text{N}(0,1) \iff \theta \sim \text{U}[0, 2\pi]$ and $R^2 \sim \text{Expo}(1/2)$ indep.
- Suggests following algorithm for generating two Gaussians $X_1, X_2 \overset{\text{i.i.d.}}{\sim} \text{N}(0,1)$:
  1. Draw angle $\theta \sim \text{U}[0, 2\pi]$ and squared radius $R^2 \sim \text{Expo}(1/2)$.
  2. Convert to Cartesian coordinates as in (1)
- From $U_1, U_2 \overset{\text{i.i.d.}}{\sim} \text{U}[0,1]$ we can generate R and θ by

$$R = \sqrt{-2\log(U_1)}, \qquad \theta = 2\pi U_2,$$

giving

$$X_1 = \sqrt{-2\log(U_1)} \cdot \cos(2\pi U_2), \qquad X_2 = \sqrt{-2\log(U_1)} \cdot \sin(2\pi U_2)$$

Rejection Sampling

❑ **Basic Idea:**

- Assume we cannot directly draw from density f.
- Tentative idea:
  1. Draw X from another density g (similar to f, easy to sample from).
  2. Only keep some of the X depending on how likely they are under f.

Rejection Sampling

Basic Idea:

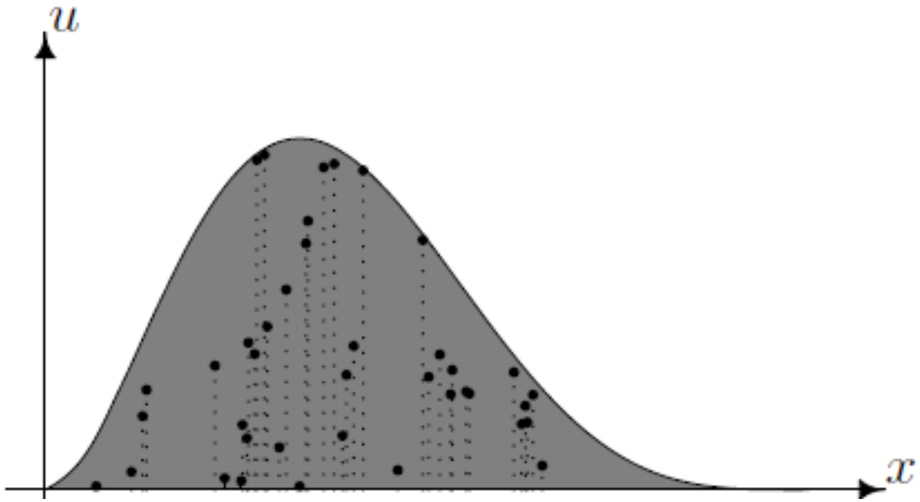
- Consider the identity

$$f(x) = \int_0^{f(x)} 1\, du = \int \underbrace{1_{0<u<f(x)}}_{=f(x,u)} du.$$

- $f(x)$ can be interpreted as the marginal density of a uniform distribution on the area under the density $f(x)$:

$$\{(x,u):\ 0 \leq u \leq f(x)\}.$$

- Sample from f by sampling from the area under the density.



Rejection Sampling

❑ Rejection Sampling Algorithm:

Algorithm 2.1: Rejection sampling

Given two densities f, g with $f(x) < M \cdot g(x)$ for all x, we can generate a sample from f by

1. Draw $X \sim g$.
2. Accept X as a sample from f with probability

$$\frac{f(X)}{M \cdot g(X)},$$

   otherwise go back to step 1.

Note: $f(x) < M \cdot g(x)$ implies that f cannot have heavier tails than g.

Rejection Sampling

❑ Rejection Sampling Algorithm:

Remark 2.1

If we know f only up to a multiplicative constant, i.e. if we only know $\pi(x)$, where $f(x) = C \cdot \pi(x)$, we can carry out rejection sampling using

$$\frac{\pi(X)}{M \cdot g(X)}$$

as probability of rejecting X, provided $\pi(x) < M \cdot g(x)$ for all x.

Can be useful in Bayesian statistics:

$$f^{\text{post}}(\theta) = \frac{f^{\text{prior}}(\theta) l(\mathbf{y}_1, \ldots, \mathbf{y}_n | \theta)}{\int_\Theta f^{\text{prior}}(\vartheta) l(\mathbf{y}_1, \ldots, \mathbf{y}_n | \vartheta)\, d\vartheta} = C \cdot f^{\text{prior}}(\theta) l(\mathbf{y}_1, \ldots, \mathbf{y}_n | \theta)$$

Rejection Sampling

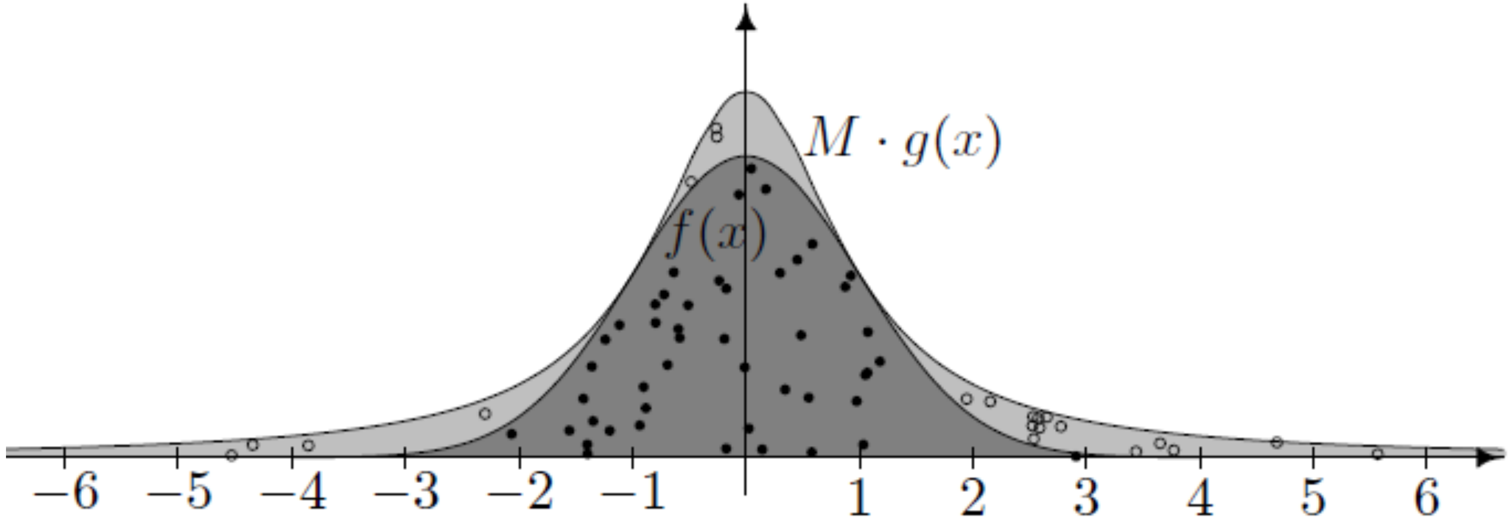
❑ **Example: Rejection Sampling from the N[0,1] distribution using the Cauchy proposal**

- Recall the following densities:

$$\text{N}(0,1) \qquad f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

$$\text{Cauchy} \qquad g(x) = \frac{1}{\pi(1+x^2)}$$

- For $M = \sqrt{2\pi} \cdot \exp(-1/2)$ we have that $f(x) \leq M g(x)$.
  $\rightsquigarrow$ We can use rejection sampling to sample from f using g as proposal.



Rejection Sampling

- **Example: Rejection Sampling from the N[0,1] distribution using the Cauchy proposal**

- **NOTE:**

  - We cannot sample from a Cauchy distribution (g) using a Gaussian (f) as instrumental distribution.
  - Whe Cauchy distribution has heavier tails than the Gaussian distribution: there is no $M \in \mathbb{R}$ such that

$$\frac{1}{\pi(1+x^2)} < M \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2}\right).$$

- **Drawbacks:**

  - We need that $f(x) < M \cdot g(x)$
  - On average we need to repeat the first step M times before we can accept a value proposed by g.

Importance Sampling

❑ Fundamental Identities:

Assume that $g(x) > 0$ for (almost) all x with $f(x) > 0$. Then for a measurable set A:

$$\mathbb{P}(X \in A) = \int_A f(x)\, dx = \int_A g(x) \underbrace{\frac{f(x)}{g(x)}}_{=:w(x)} dx = \int_A g(x) w(x)\, dx$$

For some integrable function h, assume that $g(x) > 0$ for (almost) all x with $f(x) \cdot h(x) \neq 0$

$$\mathbb{E}_f(h(X)) = \int f(x) h(x)\, dx = \int g(x) \underbrace{\frac{f(x)}{g(x)}}_{=:w(x)} h(x)\, dx$$

$$= \int g(x) w(x) h(x)\, dx = \mathbb{E}_g(w(X) \cdot h(X)),$$

Importance Sampling

- How can we make use of $\mathbb{E}_f(h(X)) = \mathbb{E}_g(w(X)\cdot h(X))$?
- Consider $X_1,\dots,X_n \sim g$ and $\mathbb{E}_g|w(X)\cdot h(X)| < +\infty$. Then

$$\frac{1}{n}\sum_{i=1}^{n} w(X_i)h(X_i) \overset{a.s.}{\underset{n\to\infty}{\longrightarrow}} \mathbb{E}_g(w(X)\cdot h(X))$$

(law of large numbers), which implies

$$\frac{1}{n}\sum_{i=1}^{n} w(X_i)h(X_i) \overset{a.s.}{\underset{n\to\infty}{\longrightarrow}} \mathbb{E}_f(h(X)).$$

- Thus we can estimate $\mu := \mathbb{E}_f(h(X))$ by
  1. Sample $X_1,\dots,X_n \sim g$
  2. $\tilde{\mu} := \frac{1}{n}\sum_{i=1}^{n} w(X_i)h(X_i)$

Importance Sampling

❑ Importance Sampling Algorithm:

Algorithm 2.1a: Importance Sampling

Choose g such that $\text{supp}(g) \supset \text{supp}(f \cdot h)$.

1. For $i = 1, \ldots, n$:
   i. Generate $X_i \sim g$.
   ii. Set $w(X_i) = \frac{f(X_i)}{g(X_i)}$.
2. Return

$$\tilde{\mu} = \frac{\sum_{i=1}^{n} w(X_i) h(X_i)}{n}$$

as an estimate of $\mathbb{E}_f(h(X))$.

- Contrary to rejection sampling, importance sampling does not yield realisations from f, but a *weighted sample* (X_i, W_i).
- The weighted sample can be used for estimating expectations $\mathbb{E}_f(h(X))$ (and thus probabilities, etc.)

Importance Sampling

❑ Importance Sampling Algorithm - Basic Properties:

- We have already seen that $\tilde{\mu}$ is consistent if $\text{supp}(g) \supset \text{supp}(f \cdot h)$ and $\mathbb{E}_g|w(X) \cdot h(X)| < +\infty$, as

$$\tilde{\mu} := \frac{1}{n}\sum_{i=1}^{n} w(X_i)h(X_i) \overset{a.s.}{\underset{n\to\infty}{\longrightarrow}} \mathbb{E}_f(h(X))$$

- The expected value of the weights is $\mathbb{E}_g(w(X)) = 1$.
- $\tilde{\mu}$ is unbiased (see theorem below)

Theorem 2.2: Bias and Variance of Importance Sampling

$$\begin{aligned} \mathbb{E}_g(\tilde{\mu}) &= \mu \\ \text{Var}_g(\tilde{\mu}) &= \frac{\text{Var}_g(w(X) \cdot h(X))}{n} \end{aligned}$$

Importance Sampling

❑ **If we know *f* up to a multiplicative constant:**

- Assume $f(x) = C\pi(x)$. Then

$$\tilde{\mu} = \frac{\sum_{i=1}^{n} w(X_i)h(X_i)}{n} = \frac{1}{n}\sum_{i=1}^{n} \frac{C\pi(X_i)}{g(X_i)} h(X_i)$$

- Idea: Estimate $1/C$ as well. Consider the estimator

$$\hat{\mu} = \frac{\sum_{i=1}^{n} w(X_i)h(X_i)}{\sum_{i=1}^{n} w(X_i)}$$

- Now we have that

$$\hat{\mu} = \frac{\sum_{i=1}^{n} w(X_i)h(X_i)}{\sum_{i=1}^{n} w(X_i)} = \frac{\sum_{i=1}^{n} \frac{\pi(X_i)}{g(X_i)} h(X_i)}{\sum_{i=1}^{n} \frac{\pi(X_i)}{g(X_i)}},$$

$\rightsquigarrow$ $\hat{\mu}$ does not depend on C

Importance Sampling

❑ Importance Sampling Algorithm - Revised:

Algorithm 2.1b: Importance Sampling using self-normalised weights

Choose g such that $\text{supp}(g) \supset \text{supp}(f \cdot h)$.

1. For $i = 1, \ldots, n$:
   i. Generate $X_i \sim g$.
   ii. Set $w(X_i) = \frac{f(X_i)}{g(X_i)}$.
2. Return

$$\hat{\mu} = \frac{\sum_{i=1}^{n} w(X_i)h(X_i)}{\sum_{i=1}^{n} w(X_i)}$$

as an estimate of $\mathbb{E}_f(h(X))$.

Importance Sampling

❑ Basic Properties of the Estimate:

- $\hat{\mu}$ is consistent as

$$\hat{\mu} = \underbrace{\frac{\sum_{i=1}^n w(X_i)h(X_i)}{n}}_{=\tilde{\mu}\longrightarrow\mathbb{E}_f(h(X))} \underbrace{\frac{n}{\sum_{i=1}^n w(X_i)}}_{\longrightarrow 1} \overset{a.s.}{\underset{n\to\infty}{\longrightarrow}} \mathbb{E}_f(h(X)),$$

(provided $\text{supp}(g) \supset \text{supp}(f \cdot h)$ and $\mathbb{E}_g|w(X)\cdot h(X)| < +\infty$)

- $\hat{\mu}$ is biased, but asymptotically unbiased (see theorem below)

Theorem 2.2: Bias and Variance (ctd.)

$$\mathbb{E}_g(\hat{\mu}) = \mu + \frac{\mu \text{Var}_g(w(X)) - \text{Cov}_g(w(X), w(X)\cdot h(X))}{n} + O(n^{-2})$$

$$\text{Var}_g(\hat{\mu}) = \frac{\text{Var}_g(w(X)\cdot h(X)) - 2\mu\text{Cov}_g(w(X), w(X)\cdot h(X))}{n} + \frac{\mu^2\text{Var}_g(w(X))}{n} + O(n^{-2})$$

Importance Sampling

❑ Finite Variance Estimators:

- Importance sampling estimate consistent for large choice of g. (only need that ...)
- More important in practice: *finite variance estimators*, i.e.

$$\mathrm{Var}(\tilde{\mu}) = \mathrm{Var}\left(\frac{\sum_{i=1}^{n} w(X_i)h(X_i)}{n}\right) < +\infty$$

- Sufficient conditions for finite variance of $\tilde{\mu}$:
  - $f(x) < M \cdot g(x)$ and $\mathrm{Var}_f(h(X)) < \infty$, or
  - E is compact, f is bounded above on E, and g is bounded below on E.
- Note: If f has heavier tails then g, then the weights will have infinite variance!

Importance Sampling

❑ **Optimal Proposal:**

Theorem 2.3: Optimal proposal

The proposal distribution g that minimises the variance of $\tilde{\mu}$ is

$$g^*(x) = \frac{|h(x)|f(x)}{\int |h(t)|f(t)\,dt}.$$

- Theorem of little practical use: the optimal proposal involves $\int |h(t)|f(t)\,dt$, which is the integral we want to estimate!
- Practical relevance of theorem 2.3:
  Choose g such that it is close to $|h(x)| \cdot f(x)$

Importance Sampling

❑ Super-efficiency of Importance Sampling:

- For the optimal g^* we have that

$$\mathrm{Var}_f\left(\frac{h(X_1)+\ldots+h(X_n)}{n}\right) > \mathrm{Var}_{g^\star}(\tilde{\mu}),$$

if h is not almost surely constant.

Superefficiency of importance sampling

The variance of the importance sampling estimate can be *less* than the variance obtained when sampling directly from the target f.

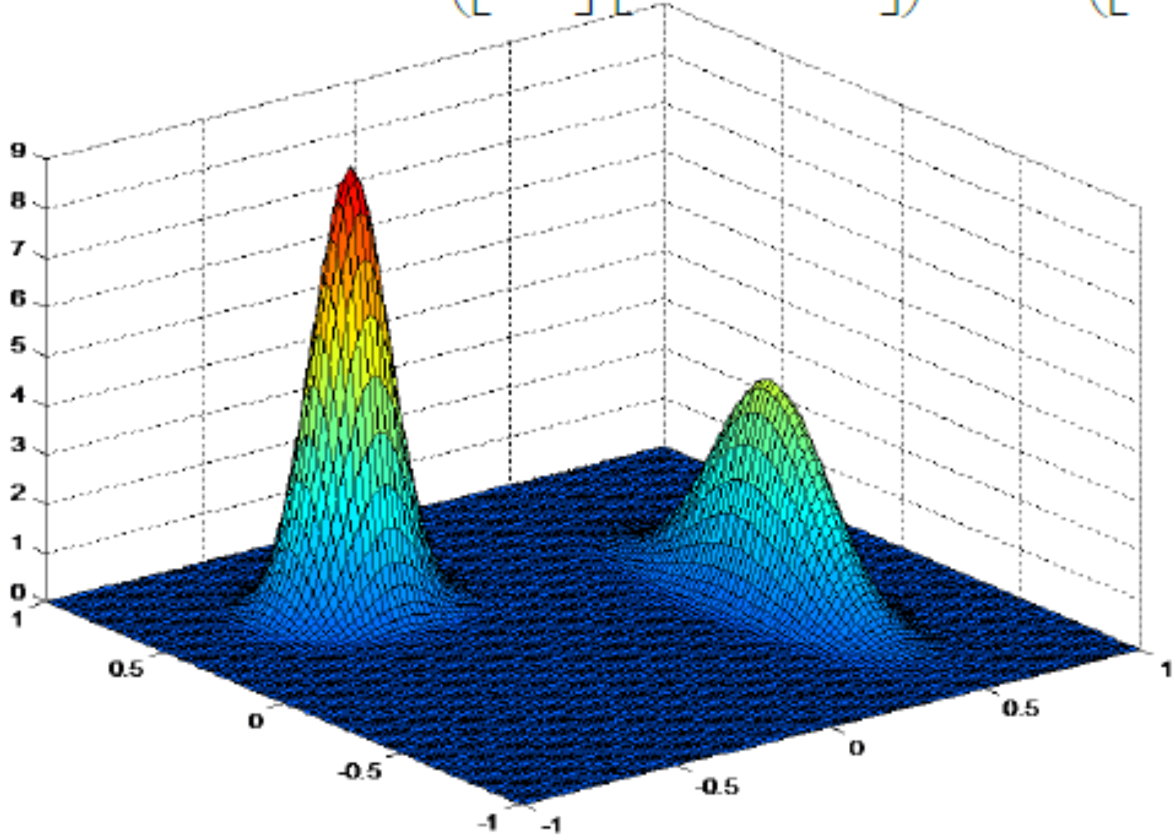
- Intuition: Importance sampling allows us to choose g such that we focus on areas which contribute most to the integral $\int h(x)f(x)\,dx$.
- Even sub-optimal proposals can be super-efficient.

Importance Sampling: Example

❑ **Calculation of integral in 2 dimensions of f(x,y):**

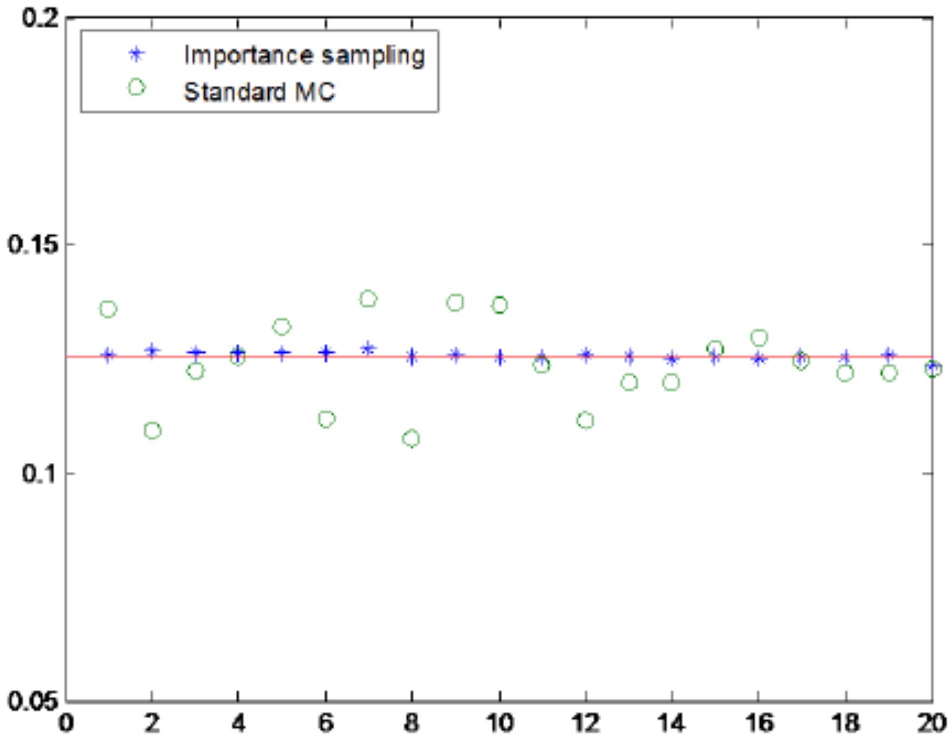
$$I = \iint_{[0,1]\times[0,1]} f(x,y)\,dxdy \qquad f(x,y) = 0.5e^{-90(x-0.5)^2-45(y+0.1)^4} + e^{-45(x+0.4)^2-60(y-0.5)^2}$$

$$\text{Proposal Distribution}: q(x,y) = 0.46N\left(\begin{bmatrix}0.5\\-0.1\end{bmatrix},\begin{bmatrix}1/180 & 0\\0 & 1/20\end{bmatrix}\right)+0.54N\left(\begin{bmatrix}-0.4\\0.5\end{bmatrix},\begin{bmatrix}1/90 & 0\\0 & 1/120\end{bmatrix}\right)$$



Importance Sampling: Example

- **Obtained Estimates:**
  - N=2000, count =20 (we take 2000 random sample points per run and run the simulation 20 times)
  - The results of importance sampling are more accurate than the standard MC method.



Bibliography

- *Monte Carlo Strategies in Scientific Computing*, J. Liu, Springer, New York, 2001.
- *Monte Carlo Statistical Methods*, C. Robert, G. Casella, Springer, New York, 2004.
- *Stochastic Methods: A Handbook for the Natural and Social Sciences*, C. Gardiner, Springer, New York, 2009.